Harvesting Indices to Grow a Controlled Vocabulary: Towards Improved Access to Historical Legal Texts

Michael Piotrowski
Law Sources Foundation
of the Swiss Lawyers Society
Zurich, Switzerland
mxp@ssrq-sds-fds.ch

Cathrin Senn
sennmantics GmbH
Thalwil, Switzerland
senn@sennmantics.com

Abstract

We describe ongoing work aiming at deriving a multilingual controlled vocabulary (German, French, Italian) from the combined subject indices from 22 volumes of a large-scale critical edition of historical documents. The controlled vocabulary is intended to support editors in assigning descriptors to new documents and to support users in retrieving documents of interest regardless of the spelling or language variety used in the documents.

1 Introduction

Until quite recently, most critical edition[1] projects produced printed books, even though the *production* of these volumes has been supported by computers since the 1960s, e.g., for concordancing, collation, and statistical analyses, as well as for bibliography management, text editing, and typesetting (see, e.g., Froger (1970)).

Modern edition projects increasingly aim to produce *digital editions* that offer linking, dynamic display of alternative readings, or the integration of related images (in particular facsimiles of original documents), audio, or video. However, the new target medium does not just offer new possibilities, but it also demands sometimes fundamental changes in the editorial process.

One affected area is indexing. In printed books, the manually constructed back-of-the-book index is the only way for readers to access the contents in a non-linear fashion. A good index is not merely a list of words occurring in the text, but it specifies concepts and introduces synonyms and, through cross-references, related terms. The possibility to perform full-text searches on digital texts therefore does not render manually constructed indices obsolete, but complements them (see Savoy (2005) for an evaluation in a comparable scenario). For editions of historical texts, a manually constructed index is indispensable, as spelling variation, meaning shifts, and multilingualism make full-text retrieval difficult for both laypersons and experts.

In book form, collective editions of shorter texts, such as letters, treaties, or charters, form one monolithic entity. The electronic medium allows for direct linking and repurposing of individual parts (or *content objects*) of a collection in new contexts, so the individual edited text is much more independent than it was in a printed volume. This has direct implications for the construction of indices: Traditionally, an index for a book is compiled when it is completed; thus, when selecting keywords, the indexer does not consider individual texts in isolation, but rather within the specific context set by the book. An indexer may thus choose one particular term for describing a concept over another one because it occurs verbatim in the majority of texts; or an indexer may choose to leave out certain possible index terms because they are self-evident in the context of the book, e.g., the index to an edition of letters is unlikely to contain the index term *letter*.

In a digital edition, in contrast, index terms should be rather thought of as metadata assigned to individual content objects to enable retrieval and reuse in different contexts. For example, if an edition of a letter is included in a thematic collection containing various types of documents, it should have the metadata information *letter*, as this may be a distinguishing feature in this collection. It also means that a collection may contain items

[1] In a narrow sense, a critical edition is a scholarly edition that tries to recover the most authentic version of a historical text from extant sources. We use the term loosely to include other types of scholarly editions, in particular diplomatic editions.

annotated by different editors, in contrast to back-of-the-book indices, which are typically created by a single indexer.

In order to ensure interoperability of index terms, a *controlled vocabulary* should be used. We define a controlled vocabulary in accordance with ANSI/NISO Z39.19-2005 (ANSI/NISO, 2005) as a set of canonical terms that are managed by an authority according to certain rules; for multiple terms referring to the same concept, a preferred term (i.e., descriptor) is defined, and a term representing various concepts is made unambiguous. A controlled vocabulary may have defined types of relationships between terms such as in a taxonomy (hierarchy), thesaurus (hierarchy, equivalence, association), or ontology (specific types of relationships like "is produced by").

Construction of controlled vocabularies is a time-consuming and labor-intensive process. Since it requires deep semantic understanding, it cannot be fully automated. However, we noted in our experiments that some stages of building a controlled vocabulary (see Shearer (2004) for a nine-step procedure to build a thesaurus) can be partially automated. In particular, we propose to harvest the information contained in subject indices from earlier or related works.

This paper describes ongoing work along these lines towards a controlled vocabulary for the *Collection of Swiss Law Sources*, a large-scale critical edition of historical texts. The vocabulary is intended to support editors in finding meaningful and agreed-upon descriptors and to facilitate retrieval of documents by both experts and laypersons. We expect that for our purposes a post-coordinate vocabulary[2] will be most useful, but the exact type and structure of the vocabulary will be defined at a later stage.

The main contributions of this paper are (1) to raise awareness for existing manually created information resources, which are potentially valuable for many tasks related to the processing of historical texts, and (2) to describe exploratory work towards using one type of resource, namely indices, for creating a controlled vocabulary.

The paper is structured as follows: Section 2 discusses related work; Section 3 gives an overview of the Collection and its subject indices; Section 4 describes the extraction of index terms and their conflation using base form reduction; Section 5 describes experiments with decompounding; in Section 6 we compare the extracted terms with the headwords of the HRG; Section 7 summarizes our findings and outlines future work.

2 Related Work

Vocabularies are inherently domain-specific. For our domain of historical legal texts, there is currently no controlled vocabulary that could be used as a basis. Despite some similarities, modern legal vocabularies such as Jurivoc[3] or the GLIN Subject Term Index[4] are not readily applicable to medieval and early modern jurisdictions (e.g., they lack concepts such as feudal tenure or witchcraft). The *Vocabulaire international de la diplomatique* (Milagros Cárcel Ortí, 1997) is an attempt at a vocabulary for describing types of historical documents, but it is not fine-grained enough and does not consider historical regional differences.

There are various approaches for automatically generating back-of-the-book indices and thus potential descriptors (e.g., Csomai and Mihalcea (2008)), but these are intended for book-length texts in a single language; in the case of historical editions, however, the documents differ widely in length, language, and age.

Romanello et al. (2009) have parsed OCR-processed *indices scriptorum* and extracted information to support the creation of a collection of fragmentary texts. Even though this is a completely different task, the approach is somewhat related to ours, in that it aims to utilize the valuable information contained in manually created indices.

3 The Collection of Swiss Law Sources

The *Collection of Swiss Law Sources* is an edition of historical legal texts created on Swiss territory from the early Middle Ages up to 1798. The Collection includes acts, decrees, and ordinances, but also indentures, administrative documents, court transcripts, and other types of documents. Since 1894, the Law Sources Foundation has edited and published more than 60,000 pages of source material and commentary in over 100 volumes.

The primary users of the Collection are historians, but it is also an important source for the Swiss-German Dictionary, which documents the

[2] See ANSI/NISO (2005) for a definition of postcoordination.

[3] `http://bger.ch/jurisdiction-jurivoc-home`

[4] `http://glin.gov/`

German language in Switzerland from the late Middle Ages to the 21st century. See Gschwend (2008) for a more detailed description of the Collection.

The primary sources are manuscripts in various regional historical forms of German, French, Italian, Rhaeto-Romanic, and Latin, which are transcribed, annotated, and commented by the editors. The critical apparatuses are in modern German, French, or Italian. Each volume contains an index of persons and places and a subject index. At the time of this writing, the Collection covers 17 of the 26 Swiss cantons to different extents.

The Collection is an ongoing project; future additions to the Collection will be created as digital editions. Instead of compiling a book, each source considered for addition to the Collection will be stored in a TEI-encoded XML document; virtual volumes, e.g., on a certain topic, place, or period, can then be created by selecting a subset of these documents. To make this possible, each document needs to contain the necessary metadata. Some of the metadata has traditionally been associated with each source text: A modern-language summary, the date, and the place of creation. In addition, each document will need to be assigned a set of descriptors.

The basis for the work described in this paper are the 22 latest volumes of the Collection, for which digital typesetting data is available; this subset is referred to as *DS21* (Höfler and Piotrowski, 2011). We have converted the typesetting files of the indices into an XML format that makes the logical structure of the indices explicit, i.e., headwords, glosses, spelling variants, page and line references, etc. The conversion process is described in detail by Piotrowski (2010).

DS21 contains volumes from ten cantons representing most linguistic and geographic regions of Switzerland and spans 1078 years. We therefore believe DS21 to be a good sample of the types of documents contained in the Collection, and we therefore expect high-frequency index terms to be good candidates for inclusion in the controlled vocabulary. The subject indices of the DS21 volumes contain a total of 70,531 entries (plus 43,264 entries in the indices of persons and places). In the work described below we have focused on the German-language volumes; the volumes in French and Italian will be considered at a later stage. The subject indices of the German-language volumes comprise a total of 47,469 entries.

weinschänckh, weinschenk, wi/ynschenck; schenckleüth *m Weinschenk* 329[12], 384[10]–386[7], 547[32]–551[9], 600[6], 601[37], 628[28], 645[21], 706[30], 740[15]–741[29], 752[8], 821[13]–824[43], 890[8]–891[13]
weinschenckhhaüßere *pl. Schenkhäuser, s.* schenckheüsser
weinstockh *m Rebstock* 665[13–18]
weinstraffen *pl. Weinbussen* 605[41]
weinter *m, s.* winter
Weintrinkverbot 313[33]–314[42], 397[21], 399[27]–400[36], 405[30]
wein umgeltner *m Umgeldverwalter* 812[10], *s.* umgelter
Weinzehnt 693[27]
Weinzins 18[16–21], 51[1]; win gůlt 396[17–22]
werber, wärber *m Söldneranwerber* 834[4–7]
werbung *f Brautwerbung* 375[2]; *Söldneranwerbung* 833[33]–834[16]
werch, wärch, werckh *n Hanf, Garn* 327[35]–328[16], 332[3], 594[35], 681[31], 825[22], 842[4]; alt w. 328[20]
werch *pl. Taten, s.* werken
werchen, wärchen, werckhen *v. arbeiten* 329[47], 350[35], 424[21], 439[27], 541[37–40], 700[7]
werchlůtten *pl. Handwerker* 178[16]
werch rybe *f Hanfreibe* 579[24]–580[21]
werd *n* 98[18]
weren, wåran, währen, wehren *v. ausrichten* 37[23], 158[6–9], 199[33], 247[13]–248[7], 350[36]–351[31], 525[23], 529[8], 664[7]; in der statt w. 99[8], 103[28 f.], 720[27]; *wehren, verwehren*

Figure 1: Extract from a subject index as it appears in a printed volume of the *Collection of Swiss Law Sources* (Rechtsquellenstiftung, 2007).

```
<p xml:id="GLO6142" class="index">
  <dfn class="hist">weinschänckh</dfn>,
  weinschenk, wi/ynschenck; schenckleüth
  <i>m Weinschenk</i> 329:12, 384:10-386:7,
  547:32-551:9, 600:6, 601:37, 628:28,
  645:21, 706:30, 740:15-741:29, 752:8,
  821:13-824:43, 890:8-891:13</p>
<p xml:id="GLO6143" class="index">
  <dfn class="hist">weinschenckhhaüßere</dfn>
  <i>pl. Schenkhäuser, s.</i>
  schenckheüsser</p>
```

Figure 2: XML version (automatically created from typesetting data) of the first two entries from Figure 1.

Figure 1 shows an excerpt of a subject index as it appears in print; Figure 2 shows two of the entries in the XML format we used as basis for the experiments described here. Since the subject indices also serve as glossaries, a particular feature is that they contain both historical and modern headwords; words in italics are modern terms, all other are historical words.

4 Extracting and Conflating Index Terms

Due to high variability of the historical index terms we decided to first concentrate on the modern index terms. Since different historians have worked on the subject indices, our first question was whether the extracted terms would overlap at all, and, if they do, to what extent and in which areas. In total, 6370 subject index word forms were extracted using a Perl script from the 16 German-language volumes. In a first step towards merging the extracted keywords, we manually removed irrelevant terms from the list of unique keywords (e.g., historical terms mistagged as modern terms), resulting in 5138 terms. We normalized the remaining entries by removing punctuation and grammatical information given with some entries. About 85% of

the unique terms occur only once. Thus, the vast majority of terms are associated with a specific volume.

Of the 15% of keywords that occur more than once the most frequent one is *Erbrecht* 'inheritance law' with 10 appearances. Although specific legal terms like *Erbrecht* are, as would be expected, relatively frequent, a similar number of keywords is linked to people's social, religious, and professional roles (reflected in terms like vagrant, baptist, pope, baker, tanner, etc.) together with terminology related to trades (for example livestock trade, animal market, sawmill). This indicates that a controlled vocabulary for the Collection should not only take into account legal terminology but also focus on roles and trades, which could potentially be covered by a separate controlled vocabulary facet (for a list of potential law subject facets see also Broughton (2010, p. 38)).

We were surprised by the small intersection between the volumes' subject indices. Looking for ways to further conflate the terms, we noted a number of mismatches due to morphological variation (such as singular and plural forms), even though subject indices are not as inflectionally rich as normal German text.

Since many index terms are highly domain-specific or specific to Swiss German (e.g., compounds of the term *Anke* 'butter' like *Ankenballen* or *Ankenhaus*), we did not use a rule-based morphological analyzer (such as GERTWOL, Stripy Zebra, or Morphisto; for an overview see Mahlow and Piotrowski (2009)) but the Baseforms tool from the ASV Toolbox (Biemann et al., 2008), which is based on pretree classifiers. The Baseforms tool does not perform morphological analysis, but is more akin to a stemmer, so that its output is not necessarily linguistically correct; however, since we are primarily interested in term conflation, this is not a major problem. When the output of the system was empty or malformed we used the original term to ensure maximum overlap. We manually reviewed and, where necessary, corrected the base forms, also to get a better understanding of the kind of potential conflations. This cut down the list of keywords from 5138 to 4881 terms, i.e., 490 terms were morphological variants that could be conflated to 233 "concepts."

The majority of term conflations concern variation in number (*Kapelle* 'chapel' and *Kapellen* 'chapels'), derivations (*Heirat* 'marriage' and *heiraten* 'to marry'), and variant compound forms (*Lehenherr* and *Lehensherr* 'liege').

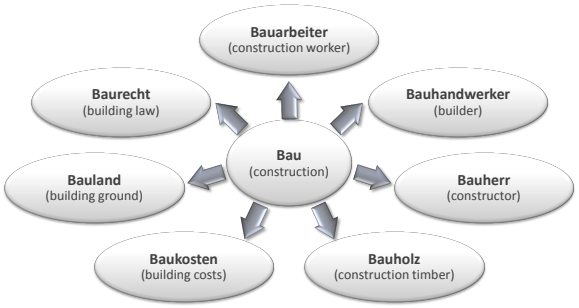

Figure 3: Map of terms based on *Bau* 'construction' with matching first compound elements.

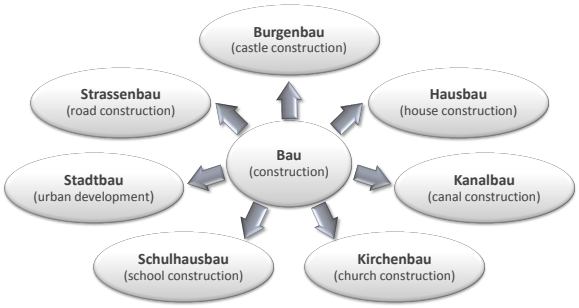

Figure 4: Map of terms based on *Bau* 'construction' with matching last compound elements.

5 Experiments with Compounds

German is well-known for its tendency to form compound nouns to express complex concepts. For vocabulary construction, compounds are interesting because related terms often share constituent parts. Our idea was therefore to use decompounding to identify potential related terms. The relationships between these terms are usually weaker than between equivalent terms (like plural and singular variants), but will still be valuable in building a controlled vocabulary. For the following experiments we used the decompounding as produced by the ASV Baseforms tool with manual corrections.

In a first experiment, we extracted groups of compound-word terms that share the same *first* element. This gives us, for example, *Bau* 'construction', *Bauarbeiter* 'construction worker', and *Bauherr* 'constructor'. The terms found in this way could, for example, be used to build a map on the topic "construction" as shown in Figure 3. In total, we found 2555 matches by first compound elements. Note that partial matching without com-

pound splitting would lead to unwanted hits like *Bauer* 'farmer' and *Baumgarten* 'tree garden'.

In a second experiment, we identified terms sharing the same *last* compound element. Overall this resulted in 2477 matches. Due to the structure of German compounds, terms sharing the final compound element are usually more closely related than those sharing the first element. Examples along the lines of *Bau* 'construction' are *Hausbau* 'house construction' and *Kirchenbau* 'church construction'; see Figure 4. Although not all of the matches will be equally relevant (for example *Erbfall* 'case of succession' and *Wasserfall* 'waterfall' are not semantically related), matches tend to point to terms on the same hierarchical level, meaning that the base form consisting of one element only (if it exists) acts as the broader term (*Bau*) of the compound matches which are the narrower terms (*Hausbau* and *Kirchenbau*).

At the moment our approach does not take into account homonyms and polysemes[5] such as *Gericht* 'court' vs. *Gericht* 'dish' or *Kirche* 'church as a building' vs. *Kirche* 'church as an institution'. Such semantic unknowns would need to be analyzed in the context of the text passages that the back-of-the-book subject indices refer to. Such a semantic review will be conducted at a later stage when the terms are prepared to be grouped in a controlled vocabulary.

6 Comparison to HRG Headwords

As noted in Section 4, the majority of index terms occur only once, i.e., in a single volume. In order to answer the question of how many of our terms are just locally useful and how many may be of more general utility, we compared our list to the list of headwords of the *Handwörterbuch zur deutschen Rechtsgeschichte* (HRG) (Cordes et al., 2008), the standard reference work on German history of law. The rationale is that the intersection of both lists contains those index terms that are highly likely to be useful as descriptors in a controlled vocabulary.

The comparison of the 3395 headwords taken from the online version of the HRG[6] (excluding entries for persons) with the 4881 stemmed index terms of our list yielded an intersection of 447 matches, i.e., 9% of our index terms also appear as headwords in the HRG.

A closer inspection shows that the rather small intersection of terms is due to the broader scope of the *Collection of Swiss Law Sources* and the fact that the HRG focuses on German rather than Swiss history. The former is illustrated by the fact that the second most frequent term in our list of index terms after *Erbrecht* is *Bäcker* 'baker', which does not appear in the list of HRG keywords. While professional roles related to legal duties like *Notar* 'notary' or *Landvogt* 'bailiff', as well as religious roles like *Papst* 'pope' or *Kleriker* 'clergyman' are also HRG headwords, terminology related to crafts and trades—like *Gerber* 'tanner' or *Schuhmacher* 'shoemaker'—is rare.

However, from a legal perspective, the terms in the intersection between the Collection and the HRG are indeed highly relevant. We also noted that high-frequency index terms from the Collection are in fact more likely to appear in the list of HRG headwords than low-frequency terms. As expected, *Erbrecht* 'inheritance law', the most frequent term in our list of index terms also occurs in the list of HRG headwords. A third of the terms appearing three times or more (306 terms) are also covered by the HRG (102 headwords), in contrast to an overlap of less than 7% for the terms occurring only once in the indices of the Collection. The index terms that occur more than once in our indices (i.e., 18% of our 4881 base form terms) account for over 46% of the terms in the intersection with the HRG headwords.

7 Conclusion and Future Work

In this paper, we have described ongoing work on the extraction of index terms from back-of-the-book subject indices in order to build a controlled vocabulary for the *Collection of Swiss Law Sources*. We have used base form reduction for term conflation and decompounding for discovering potential hierarchical relations.

We have found that index terms that are also HRG headwords are likely to be highly relevant; the terms in the intersection between our index terms and the HRG headwords will therefore be reviewed by the editors of the Collection to verify whether they are a good foundation for a controlled vocabulary.

[5] In the linguistic sense; ANSI/NISO (2005) defines homonyms and polysemes differently and would refer to homographs in this context without distinguishing whether one or more lexemes are involved.

[6] `http://www.hrgdigital.de/`

At this point, we have only examined index terms in modern language. However, the majority (85%) of modern word forms appears only once; this means that the bulk of the concepts contained in the indices must be represented by historical-language index terms. For the construction of a controlled vocabulary it is thus necessary to also consider these terms.

While there are only 6370 modern word forms (5160 unique terms) in the subject indices, we have extracted 41,099 historical word forms (28,860 unique terms). The reduction of about 30% for historical versus about 20% for modern terms indicates that historical index terms are more evenly spread across the analyzed volumes.

The percentage of historical index terms occurring only once is only slightly lower than for modern terms (80% vs. 85%); however, the historical terms exhibit a high degree of spelling variation. We therefore expect that many terms are spelling variants that can be conflated. We are currently working on methods for clustering different historical spellings of related terms.

Acknowledgements

We would like to thank Pascale Sutter for fruitful discussions and for her historical expertise.

References

ANSI/NISO. 2005. Z39.19-2005. Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies.

Chris Biemann, Uwe Quasthoff, Gerhard Heyer, and Florian Holz. 2008. ASV Toolbox: a modular collection of language exploration tools. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 1760–1767, Paris. European Language Resources Association (ELRA).

Vanda Broughton. 2010. The use and construction of thesauri for legal documentation. *Legal Information Management*, 10(01):35–42.

Albrecht Cordes, Heiner Lück, Dieter Werkmüller, and Ruth Schmidt-Wiegand, editors. 2008–. *Handwörterbuch zur deutschen Rechtsgeschichte*. Erich Schmidt, Berlin, Germany, 2nd edition.

Andras Csomai and Rada Mihalcea. 2008. Linguistically motivated features for enhanced Back-of-the-Book indexing. In *Proceedings of ACL-08: HLT*, pages 932–940, Morristown, NJ. ACL.

Jacques Froger. 1970. La critique des textes et l'ordinateur. *Vigiliae Christianae*, 24(3):210–217.

Lukas Gschwend. 2008. Rechtshistorische Grundlagenforschung: Die Sammlung Schweizerischer Rechtsquellen. *Schweizerische Zeitschrift für Geschichte*, 58(1):4–19.

Stefan Höfler and Michael Piotrowski. 2011. Building corpora for the philological study of Swiss legal texts. *Journal for Language Technology and Computational Linguistics*, 26(2):77–88.

Cerstin Mahlow and Michael Piotrowski. 2009. A target-driven evaluation of morphological components for German. In Simon Clematide, Manfred Klenner, and Martin Volk, editors, *Searching Answers – Festschrift in Honour of Michael Hess on the Occasion of his 60th Birthday*, pages 85–99. MV-Verlag, Münster, Germany.

Maria Milagros Cárcel Ortí, editor. 1997. *Vocabulaire international de la diplomatique*. Universitat de València, Valencia, Spain, second edition.

Michael Piotrowski. 2010. Document conversion for cultural heritage texts: FrameMaker to HTML revisited. In Apostolos Antonacopoulos, Michael Gormish, and Rolf Ingold, editors, *DocEng 2010: Proceedings of the 10th ACM Symposium on Document Engineering*, pages 223–226, New York, NY. ACM.

Rechtsquellenstiftung, editor. 2007. *Rechtsquellen der Stadt und Herrschaft Rapperswil*, volume SSRQ SG II/2/1: Die Rechtsquellen der Stadt und Herrschaft Rapperswil) of *Sammlung Schweizerischer Rechtsquellen*. Schwabe, Basel, Switzerland. Prepared by Pascale Sutter.

Matteo Romanello, Monica Berti, Alison Babeu, and Gregory Crane. 2009. When printed hypertexts go digital: information extraction from the parsing of indices. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia (HT '09)*, pages 357–358, New York, NY. ACM.

Jacques Savoy. 2005. Bibliographic database access using free-text and controlled vocabulary: an evaluation. *Information Processing & Management*, 41(4):873–890.

James R. Shearer. 2004. A practical exercise in building a thesaurus. *Cataloging & Classification Quarterly*, 37(3-4):35–56.